The Bell System Technical Journal
Vol. 55, No. 6, July–August 1976
Printed in U.S.A.

Evaluation of an Automatic Speaker-Verification System Over Telephone Lines

By A. E. ROSENBERG

(Manuscript received September 9, 1975)

An automatic speaker-verification system accessed by test customers from their own telephones over dialed-up lines has been evaluated. The test population consisted of over 100 male and female speakers who called up nominally once each working day over a period of five months. The operation of the system is based on a set of functions of time obtained from acoustic analysis of a fixed, sentence-long utterance. These functions are compared with stored reference functions to determine whether to accept or reject an identity claim. The system is implemented on a NOVA 800 laboratory computer. Telephone line access to the computer is via a data set hookup. Identity claims are made by keying an identification number on a Touch-Tone® *dial. Instructions and responses to the customer are made by means of a programmed voice-response system. Reference data was computed off-line and updated with the analysis data of accepted utterances. The evaluation indicated an error rate of approximately 10 percent for new customers and approximately 5 percent for adapted customers.*

I. INTRODUCTION

Speaker verification is the authentication of an individual's claimed identity by analysis of his spoken utterances. Research on an automatic system for speaker verification at Bell Laboratories has been reported in previous papers.[1–3] The system is based on an acoustic analysis of a fixed, sentence-long utterance resulting in a function of time or contour for each feature analyzed. Features selected for analysis in previous evaluations have included pitch, intensity, the first three formants, and selected predictor coefficients. The system compares the set of sample contours obtained from an unknown individual with the set of reference contours corresponding to the identity claimed by that individual. If the comparison results in an overall measure of dissimilarity which is smaller than a predetermined threshold, the identity claim is accepted. Otherwise, it is rejected.

Previous evaluations of the system have concentrated on investigating features to be analyzed and developing comparison procedures to make tbe system as effective as possible in terms of reducing overall error rate. The speech samples used in tbese evaluations were collected by wideband recording of male speakers in a sound booth. The recordings were selected and edited to eliminate botcbed utterances and nonspeech acoustic events. From the outset, however, the intent has been to provide a completely automatic system which could operate via dialed-up lines from telephones on tbe user's own premises and to include both male and female speakers in the user population. The purpose of the evaluation described in this paper was to determine bow well the system would operate under these broadened, "real-world" conditions.

Tbere are several "real-world" difficulties which are expected to be adverse to system performance. First, there are the uncontrolled and degraded environmental and transmission conditions encountered during the recording of sample utterances. Environmental conditions involve acoustic background noise and disturhances generated at the user's end and hy telephone equipment. Transmission conditions involve signal modification over dialed-up telephone lines. Telepbone transmission is nominally over a band from 300 to 3000 Hz. Tbe roll-off at 300 Hz may be gradual due to the attenuation characteristics of tbe carbon-button transmitter, or quite sharp due to tbe attenuation cbaracteristics of repeaters in some toll lines. Moreover spectral and phase distortions and variations are likely to be encountered.

The second class of problems is largely behavioral. For example, can a stable and adequate initial reference file be established based on a small numher of sample utterances collected in one sitting? Also, can botb day-to-day and long-term variations in speaking bebavior be tolerated and tracked, and can reference files be updated to reflect these cbanges in behavior?

Since the principal goal of this evaluation was to study the effect of these "real-world" conditions rather than to achieve optimum performance, the system was made more tractahle by using only pitch and intensity features for analysis.

II. SYSTEM OPERATION

Altbougb the operation of tbe system bas been described in previous papers, it will be outlined again bere, with departures pointed out, to provide a basis for discussion of tbe present evaluation.

Figure 1 provides an outline of the system operation in tbe form of a block diagram. The entire system has been implemented in software on a Data General NOVA 800 laboratory computer. Tbe two

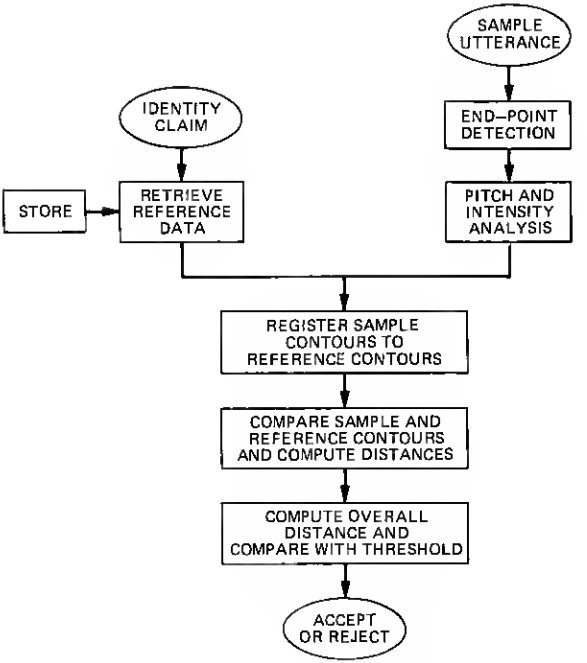

Fig. 1—System-operation block diagram.

inputs to the system are an identity claim and a sample utterance. The sample utterance used for purposes of evaluation is the all-voiced sentence "We were away a year ago." A marked interval is provided for input of the utterance. The input is subjected to 900-Hz, low-pass, analog filtering by two sections of a Rockland 1520 Dual Filter with a combined roll-off of 48 dB/octave. The filtered input is digitized at 10 kHz by means of a 12-hit analog-to-digital converter and stored on disk. The digitized input is then scanned forward from the beginning of the recording interval and hackward from the end to determine the beginning and end of the actual sample utterance. The end-point detection is accomplished hy means of an energy calculation. The delimited portion is suhjected to feature analysis which in this implementation consists of a pitch-and-intensity analysis. Pitch analysis is accomplished by means of the time-domain parallel-processing technique of Gold and Rahiner,[4] modified by Rabiner for application to telephone speech and extension to female talkers. A pitch period value is obtained every 10 ms through the course of the utterance with resolution to 100 μs, the sampling period. The resulting pitch contour is smoothed nonlinearly to bridge across unvoiced gaps and to diminish

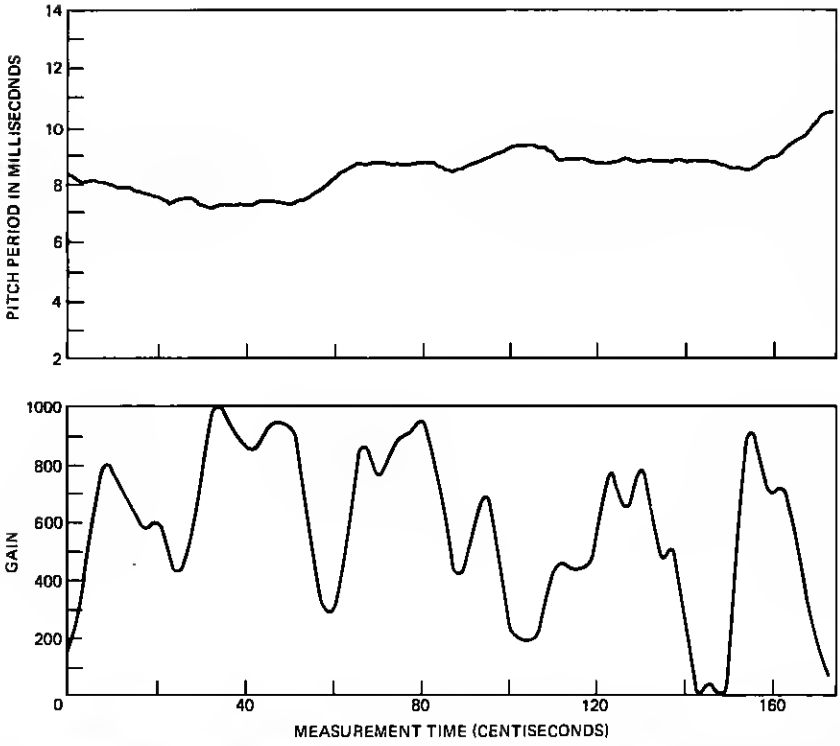

Fig. 2—Smoothed pitch period and intensity contours obtained from analysis of a sample of the test utterance.

tbe effect of singular values.* The contour is furtber subject to 16-Hz low-pass smoothing. In addition, an intensity or energy measurement is calculated every 10 ms through the course of the utterance to obtain an intensity contour. This contour is also subject to 16-Hz low-pass filtering and, in addition, is normalized to tbe peak intensity measurement resulting in a contour of relative intensity values. A typical set of pitcb-and-intensity contours is shown in Fig. 2.

Tbese contours comprise the basic patterns for verification. Tbey are compared witb a corresponding set of reference patterns associated witb tbe claimed identity. The reference patterns are obtained by averaging and combining a set of patterns obtained from sample utterances of tbe person whose identity is claimed. (Tbe referencing process bas been described in Ref. 2.) Before comparing the sample

*The pitch detector modifications, especially the nonlinear smoothing, were largely motivated by a preliminary study of the effects of telephone transmission on automatic speaker verification by O. M. M. Mitchell.[5]

and reference contours, an additional operation, time registration, is carried out. In this operation, the events of the sample contour are brought into the best possible registration with corresponding events of the reference contour by replotting the sample contour versus a modified function of time. This step is necessary to account for the normal, expected variations in speaking behavior observed in the repetition of a sample utterance by the same speaker. In previous implementations, this operation was accomplished by means of a method of steepest ascent.[1] In the present implementation, a dynamic programming technique is used. The technique is similar to those described by Sakoe and Chiba, Itakura, and Ellis.[6–8] The intensity contour is the guide contour for the procedure. The sample intensity contour is linearly stretched or compressed to the normalized length of the reference intensity contour. Then a distance is calculated between the ith point (or set of points) in the sample contour and the jth point (or set of points) in the reference contour for each i and j. The dynamic programming algorithm is used to find the path of least accumulated distances through the matrix of distances $\{d_{ij}\}$. The optimal path $i = I(j)$ $j = 1, \cdots, N$ determines the warping function required to replot the sample contour registered to the reference contour. A number of constraints are imposed so that the resulting path does not deviate excessively from the path of no warping $i = j$. The warping function obtained for the intensity contour is also applied to the pitch contour. Time registration of the intensity contour is illustrated in Fig. 3.

Following registration, the pitch and intensity contours are divided into 20 equal-length segments, as shown for intensity in the bottom panel of Fig. 3. In each segment, a set of measurements is applied to both the sample and reference contours and a squared difference is calculated specifying the dissimilarity between these contour segments for each measurement. The squared difference for each measurement and segment is weighted inversely by a variance which is calculated from the set of sample contours used to construct the reference (see Section 2.2). The effect of the variances is to weight most heavily those segments in which a particular measurement was most consistent over the set of sample contours comprising the reference. A distance for each measurement is calculated by summing the weighted squared differences over the 20 segments of a contour. In addition to four distances for each contour, based on segment-by-segment measurements, there is also a distance based on the overall cross-correlation of sample and reference contours. There are also a distance based on the cross-correlation of the pitch and intensity contours and two distances based on the amount of warping required to register the sample

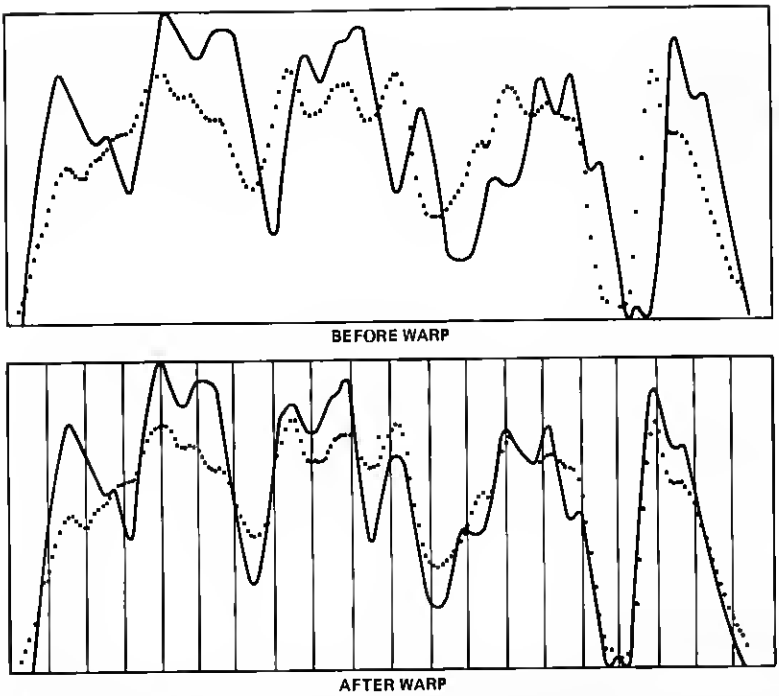

Fig. 3—Time registration of sample and reference intensity contours. The reference contour is plotted using a dotted line in both the top and bottom panel. The sample contour, solid line, is shown with its end points aligned to the reference contour, before internal registration in the top panel, and after internal registration in the bottom panel.

contours to the reference contours. The overall distance is ohtained hy a simple average over the entire set of 13 individual distances or the average over a suhset of these distances selected *a priori* for each speaker. The speaker-dependent distance-selection technique is described in Ref. 3. Finally, the overall distance is compared with a predetermined threshold to determine whether to accept or reject the identity claim.

2.1 Experimental setup and typical transaction

A block diagram of the experimental setup is shown in Fig. 4. The basic elements are the customer's phone with a *Touch-Tone* dial, a data set (Western Electric 407A), an analog-to-digital converter, and a Data General NOVA 800 computer, in which reside hoth the automatic speaker-verification programs and a programmed voice-response facility which is used to provide instructions and responses to the customer during each transaction. The voice-response system, which

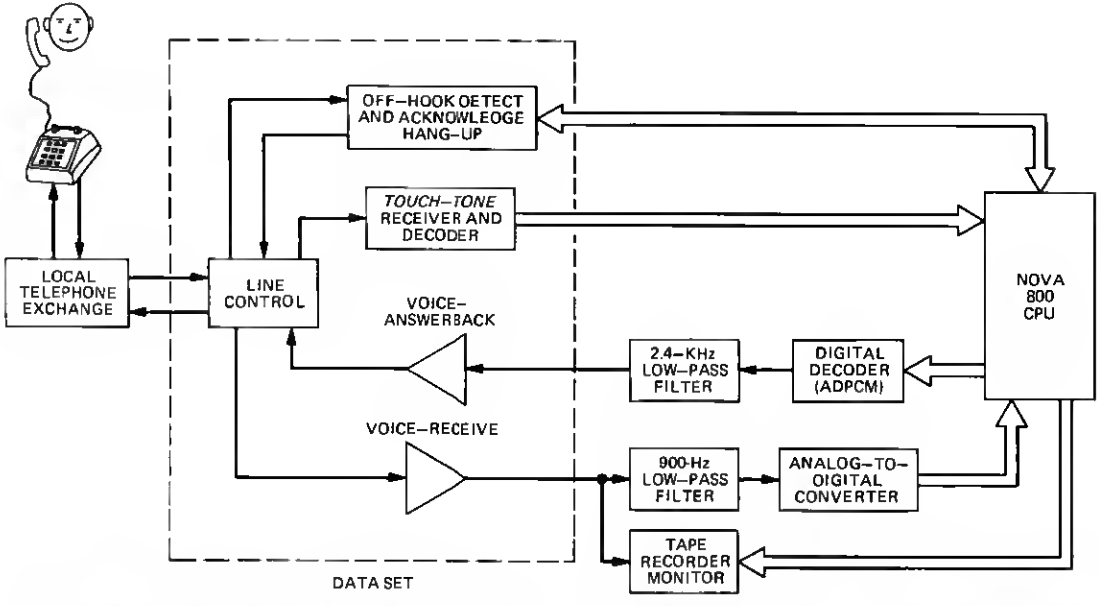

Fig. 4—Experimental setup block diagram showing the principal components of apparatus linking the customer's telephone and the computer.

uses ANPCM-coded speech message units, has been described in a paper by Rosenthal et al.[9]

A typical transaction begins with a customer dialing up the data set from his own premises via the local telephone exchange. The data set acknowledges the incoming call and provides an interrupt to the CPU, which initializes an identifying voice-response message and the instruction to the customer to dial in his identification number via the *Touch-Tone* dial on his phone. The decoded identification number constitutes the customer's identity claim. The customer is then requested to speak the test utterance within a tone-marked recording interval and the input utterance is then processed as described in the previous section. (There are several conditions that can occur during the end-point detection or pitch-and-intensity analysis indicating a bad recording that will cause the customer to be instructed to repeat the test utterance.) Following the analysis and comparison, the customer is advised of the decision to accept or reject his identity claim. A rejection, of course, constitutes a system error, a false alarm. The entire transaction from recording to verification response takes 20 to 30 seconds. A large fraction of this time is consumed by the software pitch detector. A breakdown of the computation times is shown in Table I. The major decision parameters for each transaction are appended to a log file set up for each customer.

One hundred four "customers," approximately evenly divided between adult males and females, participated actively in the evaluation by calling in nominally once each working day over a period of five months. These were all native American or Canadian speakers of English with no overt speech defects. The customers were instructed to speak the test utterance naturally and consistently from day to day.

2.2 Reference Information

The establishment and updating of reference information is an important element of the system. On the customer's initial call to the system he is requested to provide five recordings of the test utterance with approximately 10 seconds between each recording. These utterances are analyzed and the analysis data are stored in disk files. These data are used to construct the customer's initial reference file. The actual reference construction is carried out off-line during nonoperating hours and requires about 10 minutes per customer.

The following operations are included in the construction of a reference file:

(*i*) *Reference contours.* A set of reference pitch and intensity contours is constructed. The intensity contours obtained from up

Table I — Processing times for a single transaction

Operation	Time (s)
Intensity computation & end-point detection	4
Pitch analysis (3.5 s/s)	5.5 (Typical)
Time registration	3.5
Comparison & decision	2
	15
System overhead	10 (Approximate)
Total	25

to 10 customer sample files are mutually registered and averaged to provide a reference intensity contour. The corresponding sample pitch contours are mutually registered and averaged using the same warping parameters obtained for the reference intensity contour.

(*ii*) *Distance weights.* For each kind of measurement made on the contours, a variance is calculated over the sample contours used to construct the reference contours in the form

$$\sigma_j^2 = \frac{1}{N} \sum_{n=1}^{N} (s_{jn} - r_j)^2,$$

where s_{jn} is the jth measurement on the nth sample contour, r_j is the jth measurement on the reference contour, and N is the number of sample contours.

(*iii*) *Measurement selection.* The subset of the original measurement set which is most effective in separating the overall distance distribution of customer and impostor sample utterances is found. Each customer sample file and a sample file from each of 30 different customers of the same sex are used to provide the customer and impostor distributions.

(*iv*) *Threshold computation.* Estimates of equal-error thresholds for both the overall distance based on all measurements and the overall distance based on selected measurements are computed. The same sample customer and impostor files used in measurement selection are used to estimate the thresholds.

For the initial reference file, the operating threshold is set at 1.5 times the estimated equal-error threshold. This is done to compensate for the fact that the sample files used to estimate the threshold coincide exactly with the sample files used to construct the reference contours. Moreover, these sample files are obtained from utterances collected

in one session. For these reasons, the measurements and distances extracted from this set of sample files are expected to be highly correlated and are not likely to adequately reflect the expected variation over a series of independent trials. The factor of 1.5 for augmenting the threshold is somewhat arbitrary and was arrived at by inspection of the data from a preliminary experiment. The accept-reject criterion following the initial reference is based on all measurements rather than a selected set of measurements. This, again, is done because the initial sample data files do not adequately reflect the expected range of variation to provide a stable selection of measurements.

Following establishment of the initial reference, the customer calls in nominally once a day. The data files for each trial in which the customer's claim is accepted are saved. When five of these files are accumulated, the reference file is updated. The reference file is updated a second time when five additional "accepted" customer sample files are accumulated. For the fourth reference and thereafter, 10 additional customer files must accumulate. For the fourth and successive references, there are a total of 25 customer files available of which 10 are used to construct the contours and all are used to select measurements and calculate the thresholds. Following reference construction, the oldest 10 customer files are deleted, so that the maximum number of sample files per customer allowed to accumulate is 25. Also, from the fourth reference and thereafter, the operational accept/reject criterion switches over to selected measurements and the threshold is allowed to adapt. The adaptation mechanism is as follows: at each trial for which the customer claim is accepted, if the overall distance is greater than 75 percent of the current threshold, the threshold is increased by 10 percent; if the overall distance is less than 40 percent of the current threshold, the threshold is decreased by 10 percent. The reference updating procedures are summarized in Table II.

Table II — Reference updating procedures

Ref. No.	No. of Utterances Analyzed		Threshold Setting
	Total	No. Used to Construct Ref. Contours	
1	5	5	1.5 × EET* (all measurements)
2	10	10	1.5 × EET* (all measurements)
3	15	10	1.3 × EET* (all measurements)
4+	25	10	EET* (measurements selected and allowed to adapt to the next update)

* EET = Equal-error threshold.

2.3 Results

The results are described in terms of the two types of error that can occur: rejecting a (legitimate) customer claim and accepting an (improper) impostor claim. The customer-rejection rate was calculated simply from a tabulation of the number of rejections experienced by the customer population over the five-month-long period of evaluation. For each trial, overall distances were calculated both for all measurements and for selected measurements. However, the operational accept/reject criterion is hased on all measurements for the first three references and on selected measurements thereafter, as described in Section 2.2. Excluded from this tahulation were trials in which it was known that an unauthorized person used an identification number, trials in which the customer deliherately altered his utterance, trials during which the analog-to-digital recording system was operating defectively, and trials in which a faulty end-point-detection program produced misaligned analyses. (The latter two situations were quickly rectified.) Also excluded were the trials from customers whose first references failed. A first-reference failure is defined as one in which a new customer ohtained three rejections hefore obtaining the five acceptances necessary for the first update. Approximately 20 or 25 percent of the initial references failed in this way. When this occurred, the customer was requested to provide a new sample of five utterances with which to construct the initial reference. Invariably this new initial reference posed no problems.

A typical customer history is plotted in Fig. 5. Each point represents the overall distance for a particular trial. The current threshold is plotted as a broken horizontal line. An error occurs for each trial in which a point lies ahove this line, that is, where the overall distance exceeds the threshold. Three such errors occurred over the 76-trial history of this customer. The reference update in effect for each series of trials is indicated hy the numhers above the horizontal axis. The threshold was allowed to adapt trial-hy-trial following the fifth reference. The general trend for hoth distance and threshold is an initial elevation followed by a levelling off after 20 or 25 trials. This general behavior is expected and relatively easy to track. The occasional large trial-to-trial variations are not easy to track even with adaptation.

The overall results of the evaluation expressed in terms of average error rates are shown in Tahle III. Three columns of figures are shown under "reject customer." The "operational" criterion, as already mentioned, uses all measurements for the first three references and selected measurements thereafter. With the operational criterion, the average rejection rate over all customers and approximately 4500 trials is

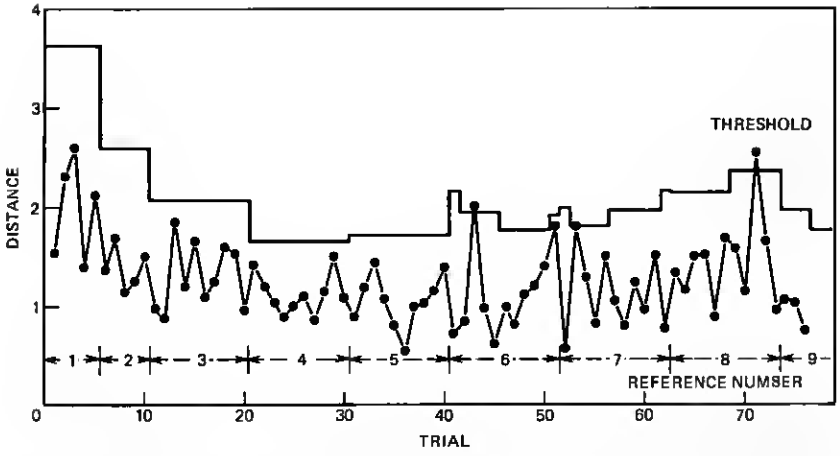

Fig. 5—Individual customer history showing the trial-by-trial overall distance throughout the period of the evaluation. Also shown is the operational threshold and reference updates through the same period.

approximately 9 percent. The median is somewhat less, approximately 7 percent. The error rates for female customers are consistently slightly higher than those for male customers. However, a statistical hypothesis test, derived by a likelihood ratio based on normal population distributions, indicated that the hypothesis of identical populations cannot he rejected at any reasonable level of significance. Under the assumption of identical populations, observed differences in error rates could be expected with a frequency as great as 20 percent. Note that the operational error rate is consistently lower than either the error rate for all measurements or for selected measurements. This is because the all-measurement criterion is preferable for early references and the selected-measurement criterion is preferahle for later ones, as shown in Fig. 6. The left half of this figure shows the average reject-

Table III — Average error rates

Customers	Reject Customer			Accept Impostor		
	Operational	All Measurements	Selected Measurements	Operational	All Measurements	Selected Measurements
Males (56)	8.20	9.46	10.65	8.78	8.91	6.06
Females (48)	9.92	10.37	12.2	10.92	10.94	7.54
Total (104)	8.99	9.88	11.33	9.72	9.79	6.70

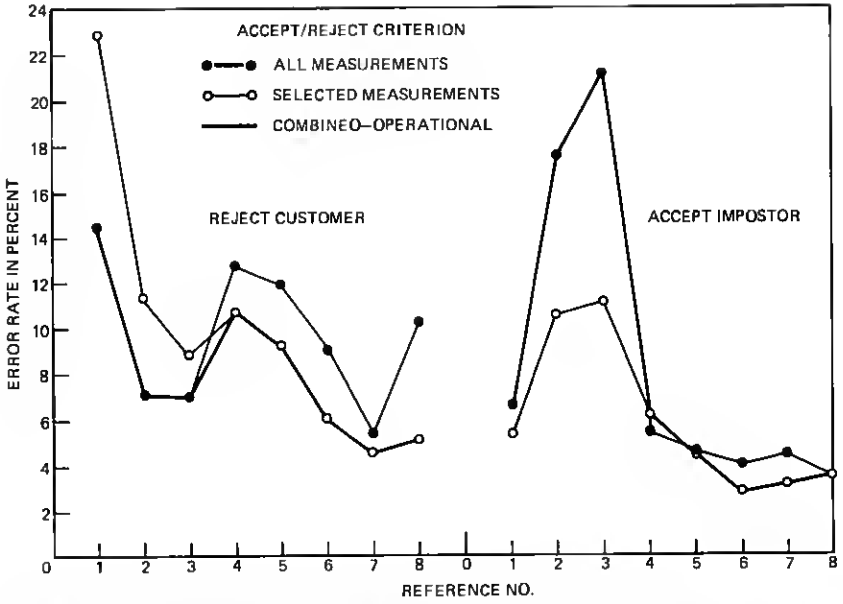

Fig. 6—Average error rates vs reference update. Reject-customer rates are plotted on the left and accept-impostor rates on the right. The thick line in each case shows the error rate under the operational criterion.

customer rate as a function of reference update. The rates for the all-measurement criterion and the selected-measurement criterion are plotted separately and the combined operational strategy is indicated by the thick line. The general trend for both criteria is the same as that observed for the history of a single customer shown in Fig. 5: a high initial error rate followed by a levelling off. It can also be seen that the operational strategy produces the minimum error rate. Overall, at early reference update stages, the reject-customer rate is of the order of 10 percent, while in later stages, the rate approaches 4 percent.

The second part of the error analysis, the tabulation of accept-impostor data, is accomplished differently. It is not practical to provide access to the system to a separate population of impostors for this purpose. Instead, a procedure was set up to systematically cross-compare selected sample utterances of each customer with the references of every other customer of the same sex. The sample file for every tenth accepted utterance of each customer is selected for this purpose, and the results tabulated. This is accomplished during the same off-line period used to update the reference files. The error rates shown in the right half of Table III are based on approximately 16,000 comparisons. The accept-impostor rates are approximately 10 percent

for the all-measurement criterion and 7 percent for selected measurements, indicating a clear advantage for the selected-measurement criterion. Unfortunately, tbe operational criterion produces no sucb improvement. The reason for this is seen in the right half of Fig. 6, where tbe accept-impostor rate is plotted versus reference number. Most of the advantage for selected measurements occurs for customer references in the early stages of updating. At these early stages, however, the operational criterion is the all-measurement one. In tbe latter stages of stable references there is no clear advantage for eitber all measurements or selected measurements. In these latter stages, as was the case for tbe reject-customer rate, tbe accept-impostor rate is approximately 4 percent.

In summary, error rates of the order of 4 or 5 percent, for botb reject-customer and accept-impostor, are obtained for customer references in advanced stages of reference updating. The overall error rates are about twice as high because of the adverse effect of errors occurring at early stages of reference updating.

Another interesting question is the effect on the error rates of varying tbe thresbold. To determine the effect of threshold variation, all the customer log files were scanned by varying the actual thresbolds in steps of 10 percent and tabulating the number of errors at eacb step by comparing the actual overall distances with the varied threshold. The results are shown in Fig. 7. All the actual thresholds are normalized to 1. Thresholds are plotted for the operational accept/reject criterion. At the normalized threshold value of 1, the error rate is approximately 9 percent, the same value shown in Table III. The plot of accept-impostor error rate as a function of threshold was obtained by comparing selected customer samples left on file at the end of the experiment with customer references. (Since the comparison data is from tbe period at tbe end of the experiment, most references were in an advanced updating stage and the error rate at the normalized threshold value of 1 is approximately 5.5 percent, considerably less than the 9.7 percent rate obtained throughout the entire period of the experiment, as shown in Table III.) It is possible to get a feeling for the amount of tradeoff obtainable when the thresbold is varied. For example, if the thresbold is set for a reject-customer rate of 4 percent, the corresponding accept-impostor rate is approximately 15 percent. Conversely, if the threshold is set for an acccpt-impostor rate of 4 percent, the corresponding reject-customer rate is approximately 11 percent.

Finally, it is of interest to survey individual error rates to get a feeling for tbe range of performance over the customer population. Histograms of individual error rates have been plotted in Fig. 8. Tbe

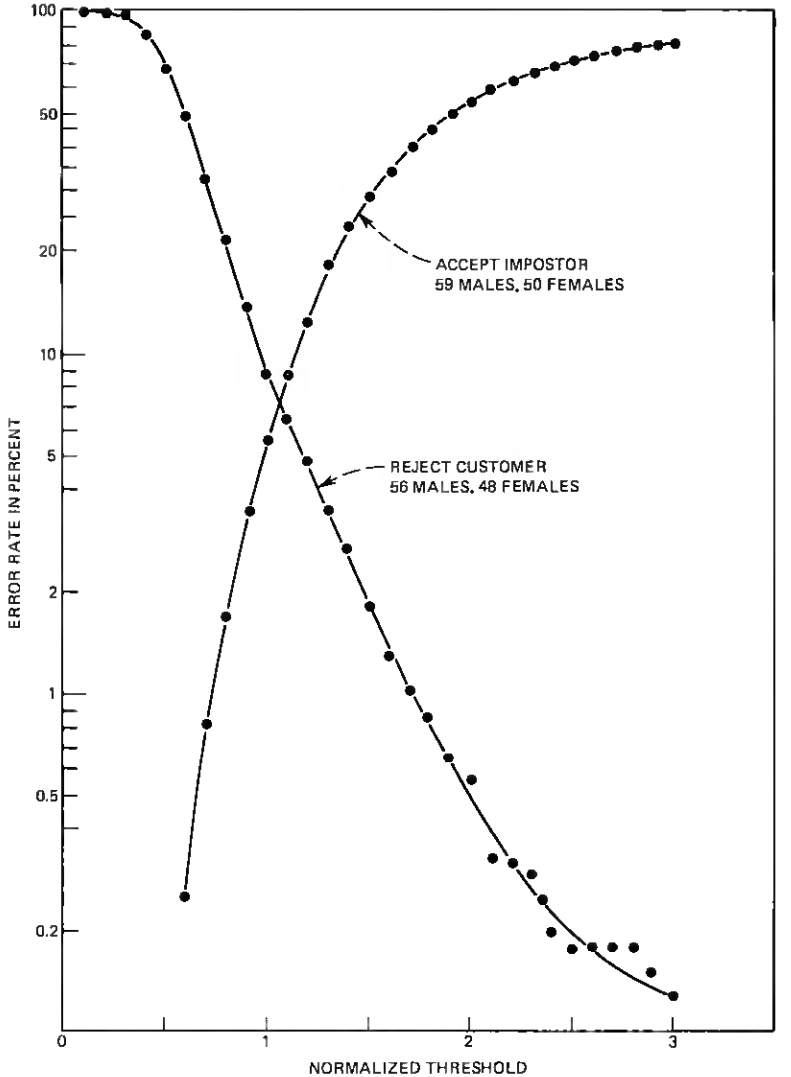

Fig. 7—Error rate vs normalized threshold. The effects are given of threshold variation on the reject-customer and accept-impostor rates using the operational accept/reject criterion.

top half shows the distribution of individual reject-customer rates under the operational criterion. About 80 percent of the customers have error rates less than 15 percent. There is, however, a small fraction of the population with excessively large rates of rejection. Some of these customers have been identified as special cases and will be

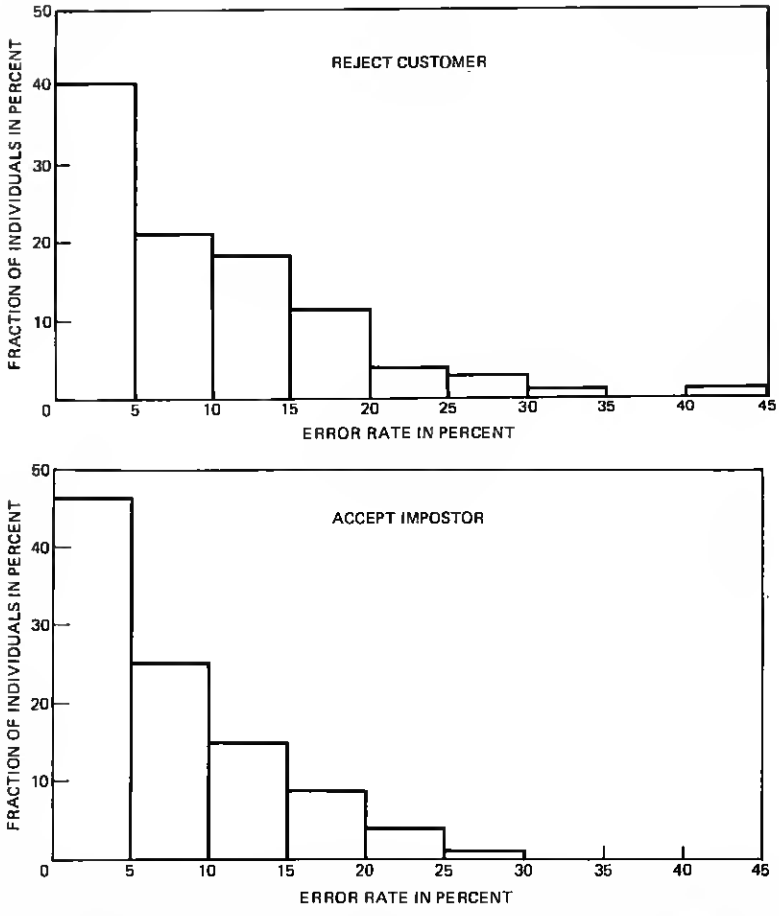

Fig. 8—Histograms showing the distribution of individual error rates over the 104-customer population. The top half shows the reject-customer distribution while the bottom half shows the accept-impostor distribution.

discussed in the next section. In the bottom half of the figure, a similar histogram is shown for the accept-impostor rates under the selected distance criterion. This distrihution is somewhat tighter, with fewer outliers than the reject-customer distribution.

III. DISCUSSION

As stated in Section I, the purpose of this evaluation was to determine how well the system could perform under "real-world" conditions. First, "real-world" conditions make it difficult for reference files to adequately represent and keep pace with normal variations in speaking behavior. To avoid encumbering the customer, the initial

referencc file is constructed from a small number of highly correlated utterances collected in one sitting. Two questions are (*i*) how adequate is this initial reference file, and (*ii*) can it be satisfactorily updated and adapted to track both trial-to-trial and long-term variations in speaking bebavior? In contrast, in previous laboratory evaluations, reference files were constructed from independent samples generally spanning a relatively long period of time witb a distinct set of test samples spanning tbe same period of time.

Second, the recordings are obtained over dialed-up lines from tbe customer's own telephone and are therefore uncontrolled and degraded in comparison with the carefully executed wideband recordings of previous laboratory evaluations.

In a previous laboratory evaluation, the equal-error rate using a pitch-and-intcnsity analysis was approximately 6 percent with tbe all-measurement criterion and 3 percent with the selected-measurement criterion.* An error rate of approximately 5 percent would be considered quite satisfactory for the "real-world" evaluation. In fact, an overall error rate of approximately 9 percent was obtained. As anticipated, the error rate varied considerahly over the course of the customers' access and reference-update history. For well-estahlished and adapted customer references, an error rate of approximately 4 percent was ohtained, which is quite acceptahle. However, the error rate of 15 or 20 percent obtained for initial customer references is unsatisfactory.

This "start-up" or initialization problem is in fact compound. In tbe first place, the number of sample utterances available during the first few reference updates is only marginally adequate to calculate reliable reference data. This is especially true for the initial reference in which the same five utterances are used to construct contours as well as to calculate weights and thresholds. It is preferable to have an independent set of utterances with which to calculate weights and thresholds. Tbe second part of the "start-up" problem is the customer's talking behavior. Morc than likely there will be large variations in talking behavior from trial to trial through the "warm-up" period of early trials until a stable or habituated talking pattern is established.

As just mentioned, the initial reference is a special case because the five samplc utterances used to calculate the reference data are collected in onc session. Generally, thc talking behavior from utterance to utterance in this session will he highly correlated. The reference data calculated from this set of utterances will tberefore he quite "tight,"

* With a predictor coefficient analysis added to the pitch-and-intensity analysis, this same evaluation yielded a 3-percent and a 1-percent error rate for all and selected measurements, respectively.

encompassing only the limited range of behavior of that session. The utterances in subsequent sessions, however, can be expected to vary considerably from those in the initial session and from eacb otber due to variations in bebavior during tbe "warm-up" period. Without compensation, then, a large customer-reject rate can be expected for tbe trials immediately following the initial reference. Tbe compensation that was attempted was to augment the calculated thresbold by 50 percent. However, it is clear tbat this was not sufficient since 20 or 25 percent of the initial references "failed," as described in the previous section, and the elevated rejection rate was 14 percent or 15 percent at this stage, as sbown in Fig. 6. Additional augmentation could well be tolerated since the accept-impostor rate at this stage is only 6 percent. Thus, the threshold could have been adjusted to yield approximately a 10-percent equal-error rate.

The situation is quite different at the first reference update, reference no. 2. At this stage, the sample utterances used to calculate tbe reference data consist of the five utterances from tbe initial session plus five more obtained in succeeding trials. The latter, as just mentioned, can be expected to vary considerably from the initial set and from each other. The result is a set of reference data which is likely to be considerably "looser" than the initial set. Consequently, tbere is a considerably reduced customer-reject rate together with a considerably increased impostor-accept rate. The same situation still bolds true at the second reference update when an additional five utterances are added to tbe reference data. In both of these reference updates, the calculated equal-error tbresbold was augmented. This is perhaps not tbe correct strategy in light of the error-rate trend.

From the third reference update, reference no. 4 and thereafter, the error rates decrease generally monotonically. It is reasonable to believe that a better choice of thresholds for tbe first three references could provide a strictly monotonically decreasing equal-error rate through the course of reference updates, say from 12 percent to 4 or 5 percent.

To provide a reduced-error rate at initial stages, a larger set of initial samples, preferably recorded at different sessions, could be obtained. Another approach which may be more practical is to provide an additional decision category, tbat of "repeat" or "defer decision." Tbus, if the distance on any trial is within a specified fraction of the threshold, additional utterances are requested until a decision can be made. This strategy is useful because tbere is a considerable probability of acceptance on a trial after rejection on a previous trial. In this experiment, a tabulation of the frequency of acceptance following rejection showed that 82 percent of tbe trials in whicb a customer was rejected were followed by trials in whicb tbe customer was accepted.

(For well-adapted customer references, reference no. 5 and thereafter, this figure increases to 91 percent.)

Anotber tabulation was carried out to assess the effect on error rate of withholding decision on trials for which $|D/T - 1| < \Delta$, wbere D and T are the overall distance and thresbold, respectively. For $\Delta = 0.05$, the customer-reject rate fell from 8.8 percent to 7.4 percent, a 15-percent improvement, with decisions withheld on 4 percent of the trials. For the same value of Δ, a 22-percent improvement was noted for the acccpt-impostor rate. For $\Delta = 0.1$, the customer-reject rate fell by 24 percent and the accept-impostor rate by 39 percent, with decisions withhcld on approximately 7 percent of the trials.

The system was shown to be capable of adapting to long-term variation in speaking behavior hy means of periodic updating of reference data and the use of an adaptive threshold. Regular and continuous usage of the system is probably necessary for sucb adaptation. Several customers had absences of two or more weeks during the course of the experiment but were not rejected with greater frequency on their return. However, the sample was too small to be conclusive. It seems likely, in fact, that many individuals will have difficulty being accepted with respect to old reference patterns after prolonged absences.

Tbe other principal "real-world" condition of interest was the effect of transmission and background noise over dialed-up telepbone lines and the background noise at the calling location. Although this was not studied in any concentrated or organized manner, generally speaking, the system seemed quite tolerant of the transmission and background conditions encountered. Over a nine-day period the following levels were monitored. The standard deviation of peak speech levels was approximately 8 dB. The background noise during recording was approximately 35 dB below the average peak speech level. Occasionally, background noise reached levels 10 dB higher than this average level. (A background noise level greater tban -18 dB with reference to average peak speech level, or a peak speech level less than -12 dB with reference to average peak speech level, would prompt a request for a new recording.) Peak speech levels for female speakers were generally 2 or 3 dB below peak levels for males, while levels for calls originating from outside the local exchange were about 3 dB helow the levels for local calls. (Approximately 25 percent of the total population called from outside the local exchange, generally via toll lines.) The only condition observed to be definitely detrimental was recording in the presence of pulse-like background noise, such as the kind originating from a typewriter or teletype console in close proximity to the telepbone transmitter. One customer who habitually called under this condition is one of the outliers in Fig. 8.

Ahnormal voice conditions were also monitored in an informal way. Mild upper respiratory infections were not ohserved to have any effect on the customer-rejection rate. However, a fairly severe case of laryngitis was ohserved for which the speaker could not provide an acceptable utterance hecause his voice "hroke" at each attempt. The most severe voice condition prohlem ohserved was diplophonia. Diplophonia is a condition associated with a husky or "raspy" voice quality. An inspection of the speech waveform for diplophonic individuals reveals voiced speech intervals in which alternate pitch periods are more strongly correlated than adjacent pitch periods. Pitch analysis for such speakers is quite difficult. Two customers, one male and one female, were ohserved to he diplophonic. The female speaker provided a rejection rate of 44 percent. It seems clear that for any system making use of pitch analysis, diplophonic speakers should he identified and, if possihle, pitch analysis modified or eliminated.

It is useful to keep in mind the results of some previous supplementary experiments in the light of the present evaluation. One experiment assessed the ahility of human listeners to perform a speaker-verification task.[10] Using the same 8-customer, 32-casual-impostor speaker set used in the earliest evaluation of the automatic system, listeners performed at an average equal-error rate of 4 percent in a series of A–B comparison trials. That is, in 4 percent of the trials in which the speakers were the same, the listeners judged them to he different, and in ahout the same fraction of the trials in which the speakers were different, the listeners judged them the same. This performance level is somewhat hetter than what has heen ohserved for this telephone evaluation, hut not as good as previous lahoratory evaluations.

Another study[11] indicated that intensively trained professional mimics have considerahly hetter chances for acceptance than casual impostors. With pitch-and-intensity analysis alone, mimic acceptance was found to he of the order of 15 percent when thresholds were set for a customer-reject rate of 1 percent. With the inclusion of predictor coefficient analysis, mimic acceptance falls sharply to a tolerable 4-percent level. Thus, pitch-and-intensity analysis by itself in a speaker-verification system may he rather vulnerahle to the efforts of determined mimics.

The practical acceptahility of a speaker-verification system depends on its expected error rate and its intended application. There are many noncritical screening applications in which the present error rate of 9 percent is acceptahle. This error rate can he improved considerahly (perhaps hy a factor of 2 or more) with the use of individual

test phrases and multiple-phrase sequential strategies. For applications with more stringent requirements, say error rates of the order of 1 percent and good protection against mimics, extended analysis techniques are necessary. As already mentioned, a laboratory evaluation which included predictor coefficient analysis satisfied these requirements.

IV. SUMMARY

The feasibility of operating the automatic speaker-verification system in the "real world" has been demonstrated. The system proved to be tolerant of many of the degraded and uncontrolled transmission and environment conditions which occur in the "real world" when customers access the system from their own premises via dialed-up telephone lines. The error rate obtained for stable and established reference patterns is approximately 5 percent, which is quite acceptable considering the abbreviated analysis used in the experiment. The greatest weakness seems to lie in the establishment of adequate initial reference patterns. It is felt that at least a partial remedy for this difficulty can result from a collection of initial samples at more than one recording session and the use of a "deferred-decision" category if the distance is within a specified tolerance of the threshold.

V. ACKNOWLEDGMENTS

A project of this complexity required the assistance of many people. The author wishes to express his appreciation to Don Bock who was responsible for providing and maintaining the system hardware, to Nancy Graham for writing the dynamic programming routines, to Carol McGonegal and Kathy Shipley for their programming assistance, to Judy Dudgeon for recruiting the "customers" and monitoring the daily experimental data, and to all the "customers" who cheerfully participated in the experiment.

REFERENCES

1. G. R. Doddington, "A Computer Method of Speaker Verification," Ph.D. Dissertation, Department of Electrical Engineering, University of Wisconsin, 1970.
2. R. C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio and Electroacoust., *AU-21* (April 1973), pp. 80–89.
3. A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker Verification," IEEE Trans. Acoustics, Speech and Signal Processing, *ASSP-23* (April 1975), pp. 169–176.
4. B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am., *46* (August 1969), pp. 442–448.
5. O. M. M. Mitchell, "Speaker Verification via Telephone," private communication.

6. H. Sakoe and S. Chiba, "A Dynamic Programming Approach to Continuous Speech Recognition," Proc. Int. Congr. Acoust., Budapest, Hungary, *3* (1971), pp. 65–68.
7. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-23* (February 1975), pp. 67–72.
8. J. H. Ellis, "Algorithm for Matching 1-Dimensional Patterns in Multidimensional Space, with Application to Automatic Speech Recognition," Electron. Lett., *5* (July 24, 1969), pp. 335–336.
9. L. H. Rosenthal, L. R. Rabiner, R. W. Schafer, P. Cummiskey, and J. L. Flanagan, "A Multiline Computer Voice Response System Utilizing ADPCM Coded Speech," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-22* (October 1974), pp. 339–356.
10. A. E. Rosenberg, "Listener Performance in Speaker Verification Tasks," IEEE Trans. Audio and Electroacoustics, *AU-21* (June 1973), pp. 221–225.
11. R. C. Lummis and A. E. Rosenberg, "Test of an Automatic Speaker Verification Method with Intensively Trained Mimics," J. Acoust. Soc. Am., *51* (January 1972), p. 131(A).